



MAPEAMENTO SISTEMÁTICO SOBRE TECNOLOGIAS PARA INTEGRAÇÃO DE BASES BIG DATA

Systematic mapping on big data base integration technologies

João Gabriel Baraldi Oliveira da Silva¹, Gabriel Silveira Dos Santos Jaques¹, Regis Rodolfo Schuch², Rodrigo Luiz Antoniazzi², Patricia Mariotto Mozzaquatro Chicon³

Resumo: Este trabalho é parte integrante de um projeto de pesquisa em desenvolvimento sobre tecnologias aplicadas a bases Big Data. Um estudo de mapeamento sistemático foi conduzido, foram examinados catorze (14) estudos relacionados ao período do ano de 2013 a 2019. Os resultados obtidos apresentam uma compreensão panorâmica da área investigada, revelando as principais tecnologias aplicadas na integração de bases Big Data. Além disso, outras informações pertinentes foram mapeadas: a distribuição das publicações ao longo do período e os autores mais ativos na área.

Palavras-chave: Big Data. Mapeamento Sistemático. Tecnologias.

Abstract: This work is part of a research project in development about technologies that are applied to Big Data bases. A systematic mapping study was conducted, fourteen (14) from 2013 to 2019 were examined. The final results showed a panoramic comprehension of the investigated area, revealing the main technologies that are applied in the integration of Big Data bases. In addition, other relevant information were mapped: the distribution of the publications throughout the period and the most active authors in this area.

Keywords: Big Data. Systematic Mapping. Technologies.

1 INTRODUÇÃO

Nos tempos atuais vivencia-se um crescente acúmulo de informações nas bases de dados das empresas. Para garantir sua permanência no mercado e competitividade, as mesmas buscam constantemente conhecimentos para fundamentar a tomada de decisão. Como muitas vezes esse conhecimento está oculto em grandes bases de dados, torna-se necessário a utilização de mecanismos computacionais que permitam a interação com os dados de forma rápida e inteligente. Com o crescimento exponencial destes dados, nos últimos anos constituiu-se o termo que conhecemos atualmente por Big Data. Neste contexto, inúmeras bases de dados possuem grande volume, alta velocidade de crescimento e grande variedade

¹ Discentes do curso de Ciência da Computação, da Universidade de Cruz Alta - Unicruz, Cruz Alta, Brasil. E-mail: drplaguejg@gmail.com, gabrielnjaques@gmail.com

² Docentes do Curso de Ciência da Computação, da Universidade de Cruz Alta - Unicruz, Cruz Alta, Brasil. E-mail: rschuch@unicruz.edu.br, rantoniazzi@unicruz.edu.br

³ Pesquisadora do Grupo de Pesquisa ENGETEC, Docente do Curso de Ciência da Computação, da Universidade de Cruz Alta - Unicruz, Cruz Alta, Brasil. E-mail: pmozzaquatro@unicruz.edu.br



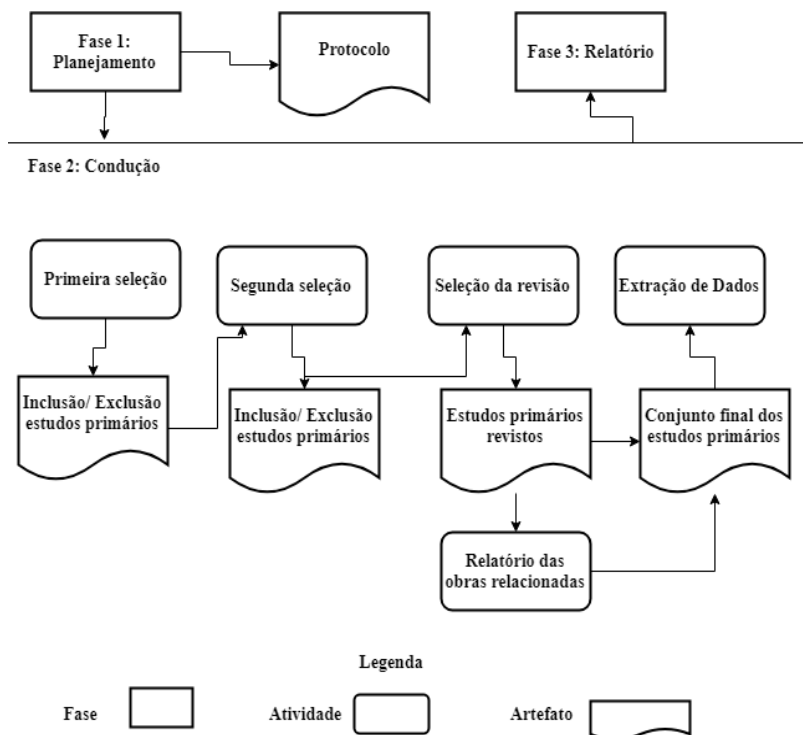
(LANEY, 2001). Entretanto, necessita-se buscar novas formas para analisar e processar esses grandes volumes de dados, que muitas vezes estão desestruturados, não podendo ser analisados através de técnicas tradicionais para extração de informações e geração de conhecimentos.

Neste contexto, no presente trabalho é realizado um mapeamento sistemático sobre as tecnologias que estão sendo empregadas na *Extract Transform Load (ETL)*, bem como as ferramentas para a manipulação e integração dos dados no contexto de Big Data.

2 MAPEAMENTO SISTEMÁTICO

Estudos de mapeamento sistemático são projetados para dar uma visão geral de uma área de pesquisa através da classificação e contando contribuições em relação às categorias dessa classificação (Petersen et al., 2008), (Kitchenham, 2007). A condução deste mapeamento sistemático baseia-se em um protocolo pré-definido que considera os *guidelines* definidos por Petersen et al. (2008). Na Figura 1, ilustra-se o processo de condução do mapeamento sistemático.

Figura 1- Processo adotado na condução do Mapeamento Sistemático



Fonte: Elaborado pelo Autor



2.1 Planejamento da pesquisa e Condução dos estudos primários

A pesquisa foi desenvolvida nas seguintes etapas: (i) definição de questões de pesquisa, (ii) realização de buscas por estudos primários relevantes, (iii) triagem dos documentos, (iv) *keywords* dos resumos, e (v) extração e análise dos dados.

As questões de pesquisa foram as seguintes:

QP1: Qual a distribuição temporal dos trabalhos no contexto do Big Data?

QP2: Quais tecnologias estão sendo aplicadas na ETL, no contexto de Big Data?

QP3: Quais ferramentas estão sendo utilizadas para a manipulação dos dados de Big Data?

QP4: Quais linguagens estão sendo utilizadas para a integração dos dados de Big Data?

QP5: Qual foi o resultado das pesquisas?

As buscas foram realizadas nas seguintes bases de dados: *IEEE Xplore Digital Library*⁴ e *Google Scholar*⁵. Para conduzir a pesquisa, foram utilizadas as seguintes palavras chave na língua portuguesa “*tecnologias, Big Data, ferramentas, manipulação dos dados, linguagens e integração dos dados*”. As mesmas palavras foram também pesquisadas na língua inglesa: “*technologies, Big Data, tools, data manipulation, languages e data integration*”.

A *string* de busca foi definida com base nas palavras – chave. A seguir é apresentada a *string* de busca: ((*technologies OR tools*) AND (“*BIG DATA*” OR “*prossesing data*” OR “*data science*”)) AND (“*data integration*” OR *ETL*) AND (*language*) AND (*database*)).

Após criar a *string* de busca, foi definido o tempo para coletar as pesquisas, ou seja, as buscas foram realizadas no período do ano de 2013 ao ano de 2019.

Com a *string* de busca e o período definidos, criou-se os critérios de seleção: critérios de inclusão e exclusão e qualidade.

Como critérios de inclusão cita-se: pesquisas em Inglês, pesquisas em Português e pesquisas com relevância que apresentam os autores da obra.

Como critérios de exclusão: pesquisas que não apresentam as tecnologias estão sendo aplicadas no ETL, pesquisas que não definem quais ferramentas estão sendo utilizadas para a manipulação dos dados no contexto de big data, pesquisas que não citam as linguagens as quais estão sendo utilizadas para a integração dos dados no contexto de big data e pesquisas publicadas anteriormente ao ano de 2014.

⁴ ieeexplore.ieee.org

⁵ scholar.google.com.br



Como critérios de qualidade cita-se: pesquisas que apresentam uma definição clara dos seus objetivos, pesquisas com resultados integrais/completas e pesquisas com propostas distintas.

3 RESULTADO

Esta seção apresenta o relatório de resultados e síntese dos dados. A Tabela 1 ilustra o resultado do mapeamento sistemático.

Tabela 1 – Resultado do Mapeamento Sistemático

Estudo 1: Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration	
Autor	Srividya K Bansal
Ano	2014
Tecnologias	ETL, UML, diagramas, linked data, IBMM, LOD, online association rule mining, workflow, URI, RDF, OWL, Protegé Ontology Editor
Ferramentas - softwares	GIS, RDF, semantic ETL Framework, IBM Infosphere, Oracle Warehouse Builder, Microsoft SQL Server Integration Services, Informatica Powercenter for Enterprise Data Integration, Pentaho Kettle, Talend Open Studio, Oxygen XML editor, Apache Jena
Linguagem	LDL, SPARQL, RDF, OWL
Estudo 2: Heterogeneous Data Security Fusion System Based on Binary Tree	
Autor	Tao Lv, Cai Fu, Deliang Xu, Guohui Li, Lansheng han
Ano	2017
Tecnologias	Federation(Data Warehouse, Data Service, Data Grid), Pegasus, SIMS, Object-Based-Mapping, DBLP
Ferramentas - softwares	Heterogeneous Data Security Fusion System Based on Binary Tree
Linguagem	XML
Estudo 3: Business Intelligence Revisited	
Autor	Jianwen Su, Yang Tang
Ano	2017
Tecnologias	ETL, OLAP, BPM, BI, mineração de processos, workflow, Data agnostic models, UML, Data aware models, BPEL, YAWL, Storage aware models, jBPM, Artifact-centric models, DEZ-Flow, GSM, Semantic Process Models, SeGa, Activiti, WSDL, WSCDL
Ferramentas - softwares	Java Hibernate, Oracle DBMS
Linguagem	SQL, LTL, Workflow log, XQuery
Estudo 4: The Implementation of a Data-Accessing Platform Built from	



Big Data Warehouse of Electric Loads	
Autor	Sheng-Cang Chou, Chao-Tung Yang, Fuu-Cheng Jiang, Chih-Hung Chang
Ano	2018
Tecnologias	Big data warehouse, ETL, HDFS, NoSQL, DataNodes, RDBMS, HBase, RDB, Workflow, metastore, Spark cluster, OLAP, Experimental Results, System Architecture, System Implementation
Ferramentas - softwares	Impala SQL, Apache HiveQL, Apache Spark SQL, Apache Hadoop, MapReduce, Apache Sqoop, Cloudera, PostgreSQL Database, MADlib, Hadoop, YARN, Apache Mesos, Cassandra, OpenStack Swift, Amazon S3, JDBC, MySQL, Solr, Hue
Linguagem	SQL, HiveQL
Estudo 5: A Study on Big Data Integration with Data Warehouse	
Autor	T.K. Das, Arati Mohapatro
Ano	2014
Tecnologias	MapReduce, Hadoop Distributed File System, HBase Database, Apache Hadoop, MAD System, Clusters, metadata, Map Phase, Raw Data, Storing Data (Dimensional, Normalized), Data Warehouse, ETL, MPP/Shared nothing algorithms, LOBs, ELT, Schema Less storage, Data node, Name node, Hive, Cheetah, Ad-hoc, Social Master Data Management, Commercial BI Tools
Ferramentas - softwares	Hadoop, Hive, HBase Database, MapReduce, Cheetah, Pig, Informatica 9.5, Cognos, Microstrategy
Linguagem	NOSQL, Java, SQL, XML, Txt, HiveQL, Scope
Estudo 6: Mapping Abstract Queries to Big Data Web Resources for On-the-fly Data Integration and Information Retrieval	
Autor	Hasan M. Jamil
Ano	2014
Tecnologias	NoSQL, Queries, BioFlow, BioBike, Query Synthenizer, AutoStruct, Data Model, meta-data, concept graph, NFNF relation schemes, Meta-language, Graph Matching based Query transformation, ER Sense, AutoPipe, Big Data Web Resources
Ferramentas	LifeDB, BioVis, PDB, MySQL, eXist



- softwares	
Linguagem	SQL/XQuery, Perl
Estudo 7: Integration of Big Data for Querying CAN Bus Data from Connected Car	
Autor	Lionel Nkenyereye, Jong-Wook Jang
Ano	2017
Tecnologias	CAN Bus Data, Remote On-line Vehicle Diagnostics System (ROVDS), Diagnostic Troubles Codes (DTC), CVTA, ECU, Hadoop Clusters, HDFS
Ferramentas - softwares	MapReduce, Hadoop, Apache Hive, Sqoop, MySQL, MongoDB, SQLite, EC2
Linguagem	HiveQL, CVS, SQL
Estudo 8: Visualization and Analytics of Biological Data by Using Different Tools and Techniques	
Autor	Raiha Tallat, Rana M. Amir Latif
Ano	2019
Tecnologias	GAP, Data mining, Graph analytics, Labeled property graph, Motif Detection, Motifs, Data model, Access Methods and Query Languages, Database underlying Storage, Caching, Concurrency, Scalability, path analytics, Centrality Analytics, Workflow, Degree Centrality, Histograms, Big Data Graph Analytics
Ferramentas - softwares	GAP, Titan, Sparksee, Neo4j, InfiniteGraph, GraphBase, HypergraphDB, InfoGrid, FlockDB, VelocityGraph, CIRCOS, FANMOD, Circa
Linguagem	HTML, NoSQL
Estudo 9: Big Data: A Review	
Autor	Seref Sagiroglu, Duygu Sinanc
Ano	2013
Tecnologias	Volume, Velocity, Variety, Big Data Samples, Big Data Topics (Healthcare, Public sector, Retail, Manufacturing, Personal location data), MapReduce, Big Data Methods, Map Step, Reduce Step, ETL, JobTracker, HDFS, Workflow, Hadoop Clusters, HPCC Systems (Data Refinery(Thor), Data Delivery Engine(Roxie), ECL), Data Warehouse, DFS, KDD, Data mining, Visual analysis, Statistical Analysis
Ferramentas - softwares	Lucene, Solr, Hadoop, HBase, Cassandra, Hive, Pig, Voldemort, Storm, IndexTank, MapReduce, BigTable, Avro, Oozie, Chukwa, Flume,



	ZooKeeper, GreenPlum, RDBMS, Thor, Roxie
Linguagem	Java, SQL, ECL
Estudo 10: Big data technologies and Management: What conceptual modeling can do	
Autor	Veda C. Storey, Il-Yeol Song
Ano	2017
Tecnologias	The Vs of Big Data, Decision making, Business Insights, Process Optimization, EDW, Cloud-based service, dark data, open data, external data, Data Virtualization, ETL, ELT, ROI, SW Solution, machine learning, statistics visualization, optimization, data mining, 7Vs(validity, volatility), Data stack, Big data ecosystem, Enterprise information management, Data science, prescription techniques, predictive analytics, descriptive techniques, Data-driven paradigm, Data speaking, Datafication, Granular data, Codd's Relational Model, RDMS, Clusters, SPOF, ACID, BASE, Data storage, Data processing, Data access, Management, Workflow, Data serialization, DAG, ETL, Streaming, complex analytics, DataFrames, RDDs, Schema-less, OLTP workloads, IMDG, IMDBs, In-memory computing, Hybrid cloud computing, SaaS, PaaS, IaaS, iPaaS, OLAP, dashboards, Schema-on-read, Dynamic Column Generation, Linking, schema-on-write, database modeling, Key-value store, Column Store, Query, Document Store, Graph Database, GIBIR, Assessment, OLTP, CRISP-DM, CDO, Data analytics lifecycle, model planning, CAP
Ferramentas - softwares	NoSQL, Amazon web service, HDFS, Hadoop, DB2, Cassandra, NewSQL, PostgreSQL, Sybase IQ, SAP HANA, Microsoft SQL Server, MapReduce, Oracle, Teradata, Neo4j, Dynamo Paper, Mongo DB, VoltDB, Redis, BigTable Paper, Oozie, Hive, Pig, Mahout, Avro, Sqoop, HBase, Chukwa, Flume, ZooKeeper, Spark, Spark Streaming, MLlib, GraphX, Tachyon, MemSQL, NuoDB, Clustrix, Google spanner, gartner.com, Oracle NoSQL, academy.datastax.com, CouchDB, Titan, OrientDB, FoundationDB, Watson
Linguagem	SQL, XML, Scala, Python, R, JSON, BSON
Estudo 11: Distributed Service Orchestration: Eventually Consistent Cloud Operation and	



Integration	
Autor	Merlijn Sebrechts, Thomas Vanhove, Gregory Van Seghbroeck, Tim Wauters, Bruno Volckaert, Filip De Turck
Ano	2016
Tecnologias	Orchestration, DevOps, Omega, Mesos, meta-schedulers, OASIS, ETL, integration, cluster, FIFO queue, Hybrid integration, PaaS, SaaS
Ferramentas - softwares	Apache Hadoop, TOSCA, Ubuntu Juj
Linguagem	TOSCA, SQL, Java
Estudo 12: Linked Enterprise Data Model and its use in Real Time Analytics and Context-Driven Data Discovery	
Autor	Kunal Taneja, Qian Zhu, Desmond Duggan, Teresa Tung
Ano	2015
Tecnologias	NoSQL, Linked Open Data, ETL, Lambda architecture, Linked Enterprise Data Model, Smart Water Use Case, RDF, Sensor Reading Normalization, Sensor Reading Projection, Event Direction, Streamed and Historical data, HTTP, Url, Extensibility and inclusion, Graph based representation, Queryability, REST APIs, OSI Pi Historian, Cluster, EC2 instances, Serialization, Batch Interval, SOLID Architecture, Linked Open Data Cloud, Linked Sensor Data
Ferramentas - softwares	HDFS, Redshift, Talend, AWS, Apache Spark Streaming, AWS Dynamo(DynamoDB), Cassandra, AWS S3, Amazon Kinesis, Metanautix, Kryo Serialization, MongoDB, SAP
Linguagem	Java, UML, DDL, SPARQL, TOSCA
Estudo 13: Raising Time Awareness in Model-Driven Engineering	
Autor	Amine Benelallam, Thomas Hartmann, Ludovic Mouline, Francois Fouquet, Johann Bourcier, Olivier Barais, Yves Le Traon
Ano	2017
Tecnologias	MDE, Time awareness, existing graph processing, temporal query languages, temporal queries, temporal pattern matching, data-driven application, xDSLMS E-Motions, DSL, OCL, Smart Grid, micro-grids, time granularity, temporal relational databases, Time-aware modeling, Temporal graphs, iterative graph computation, KMF, ER, Mape-K, IoT, CPS, UID, CDO, EMFStore, NoSQL, UML2Relational mapping,



	metamodel, MADS, Portal
Ferramentas - softwares	Chronos, ImmortalGraph, HGS, Kineograph, GraphTau, CloudMF, ThingML, MindCPS, NoSQL, NEOEMF, GreyCat, TGraph
Linguagem	UML, SQL, TSQL2, Cypher, TEG-QL
Estudo 14: Research and Implementation of Data Fusion Method Based on RDF	
Autor	Bin Yu, Chen Zhang, Yitong Li, Jiangyan Sun
Ano	2018
Tecnologias	Data fusion, RDF, Semantic Web, C3I, URI, Table mapping, Field mapping, Graph, Flow chart
Ferramentas - softwares	Jena
Linguagem	XML, SPARQL, R2RML, SQL

Fonte: Elaborado pelo Autor

A Figura 2 ilustra a quantidade de publicações por ano. Constatou-se que no ano de 2017 existiram mais pesquisas relacionadas a temática abordada neste artigo.

Figura 2 – Quantidade de publicações

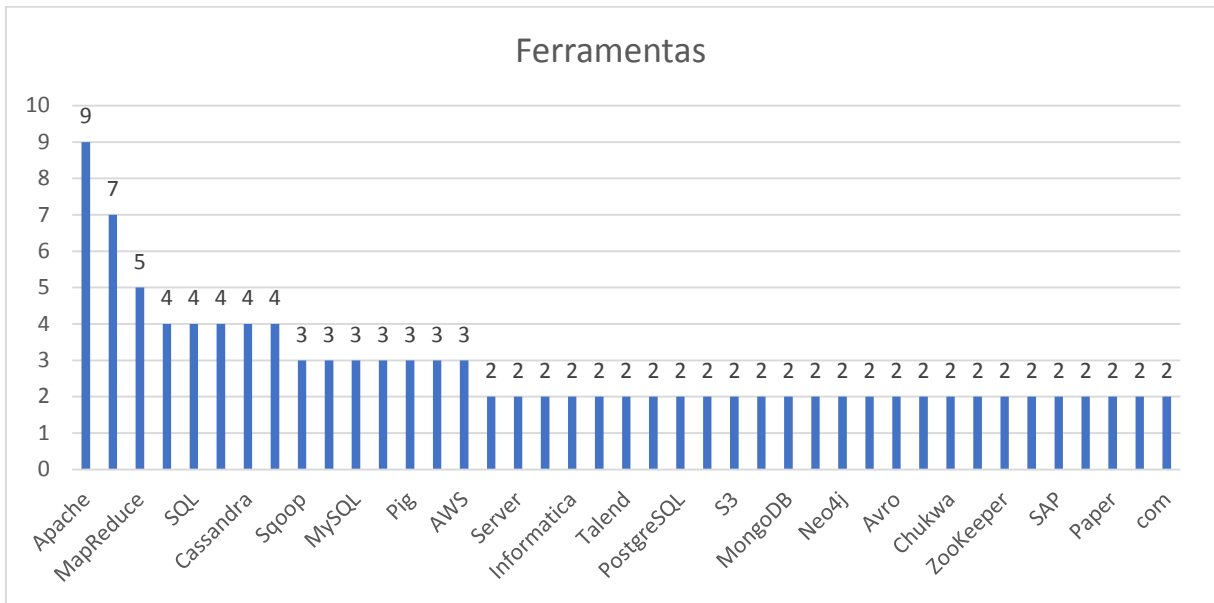


Fonte: Elaborado pelo Autor

A Figura 3 ilustra as ferramentas utilizadas para a manipulação dos dados no contexto de big data. Constatou-se que as ferramentas Apache e MapReduce foram as mais citadas nos estudos analisados.



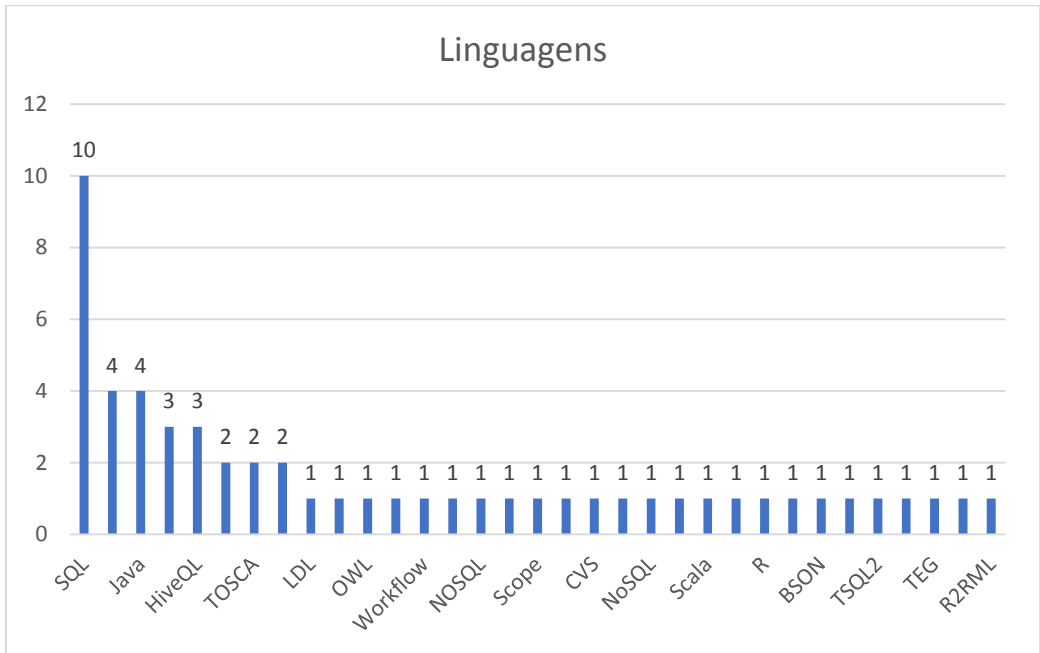
Figura 3 – Ferramentas para manipulação de dados



Fonte: Elaborado pelo Autor

A Figura 4 apresenta as linguagens mais utilizadas para a integração dos dados no contexto de big data. Constatou-se que SQL, XML e Java foram as mais utilizadas.

Figura 4 – Linguagens para a integração dos dados



Fonte: Elaborado pelo Autor

4 CONSIDERAÇÕES FINAIS

As tecnologias de Big Data são essenciais para as empresas e precisam ser bastante estudadas para o avanço do fluxo dos dados e do controle dos mesmos, com isso, é preciso de



muita estatística e uma integração precisa desses dados. Desse modo, o estudo e aplicação das técnicas de Big Data são de suma importância e para o mercado atual, devendo serem aplicadas com cautela.

O presente estudo auxiliou na análise de ferramentas para a integração de dados no contexto de Big Data, apresentando as principais publicações e autores relevantes e ativos na área. O trabalho é parte integrante de um projeto de pesquisa em desenvolvimento que tem por objetivo desenvolver um framework para a integração e dados heterogêneos provindos de diversas fontes.

REFERÊNCIAS

BANSAL, Srividya K. **Towards a Semantic Extract-Transform-Load (ETL) framework for Big Data Integration**. IEEE International Congress on Big Data. Dept. of Engineering & Computing Systems Arizona State University. DOI 10.1109/BigData.Congress, 2014.

BENELALLAM, Amine; HARTMANN, Thomas; MOULINE, Ludovic; FOUQUET, Francois; BOURCIER, Johann; BARAIS, Olivier; LE TRAON, Yves. **Raising Time Awareness in Model-Driven Engineering**. ACM/IEEE 20th International Conference on Model Driven Engineering Languages and Systems. University of Rennes 1, IRISA, INRIA Centre Rennes; Interdisciplinary Center for Security Reliability and Trust, University of Luxembourg. DOI 10.1109/MODELS.2017.

CHOU, Sheng-Cang.; YANG, Chao-Tung.; JIANG Fuu-Cheng.; CHANG, Chih-Hung. **The Implementation of a Data-Accessing Platform Built from Big Data Warehouse of Electric Loads**. 42nd IEEE International Conference on Computer Software & Applications. Department of Computer Science Tunghai University.; College of Computing and Informatics. 2018.

DAS, T. K.; MOHAPATRO, Arati. **A Study on Big Data Integration with Data Warehouse**. *International Journal of Computer Trends and Technology*. School of Information Technology & Engineering, VIT University; Department of Computer Science, Bangalore City College, 2014.

JAMIL, Hasan M. **Mapping Abstract Queries to Big Data Web Resources for On-the-fly Data Integration and Information Retrieval**. ICDE Workshops 2014. Department of Computer Science University of Idaho, 2014.

KITCHENHAM, B.; Charters, S. **Guidelines for Performing Systematic Literature Reviews in Software Engineering**, Tech. rep., Technical report, EBSE Technical Report EBSE-2007-01, 2007

LANEY, D. **3d data management: Controlling data volume, velocity and variety**. Application Delivery Strategies, META Group, v. 949, 2001.

LV, Tao.; FU, Cai.; XU, Deliang.; LI Guohui.; HAN, Lansheng. **Heterogeneous Data Security Fusion System Based on Binary Tree**. IEEE 2nd International Conference on Big



Data Analysis. School of Computer, Huazhong University of Science and Technology modeling can do. J. Mack Robinson College of Business, Dept. of Computer Information Systems, Georgia State University; College of Computing & Informatics, Drexel University. *Editora Elsevier, 2017.*

NKENYEREYE, Lionel; JANG, Jong-Wook. **Integration of Big Data for Querying CAN Bus Data from Connected Car.** ICUFN 2017. Department of Computer Engineering Dong-Eui University, 2017.

PETERSEN, K.; Feldt, R.; Mujtaba, S. ; Mattsson, M. **Systematic mapping studies in software engineering**, in: 12th International Conference on Evaluation and Assessment in Software Engineering, vol. 17, 2008, p. 1. Providence University. DOI 10.1109/COMPSAC.2018.

SAGIROGLU, Seref; SINANC, Duygu. **Big Data: A Review.** Gazi University Department of Computer Engineering, Faculty of Engineering Ankara, Turkey. 978-1-4673-6404-1/13/\$31.00 ©2013 IEEE

SEBRECHTS, Merlijin; VANHOVE, Thomas; VAN SEGHBROECK, Gregory; WAUTERS, Tim; VOLCKAERT, Bruno; DE TURCK, Filip. **Distributed Service Orchestration: Eventually Consistent Cloud Operation and Integration.** IEEE International Conference on Mobile Services. Department of Information Technology, Ghent University. DOI 10.1109/MS.2016

STOREY, Veda C.; SONG Il-Yeol. **Big data technologies and Management: What conceptual**

SU, Jian.; TANG, Yan. **Business Intelligence Revisited. Fifth International Conference on Advanced Cloud and Big Data.** Department of Computer Science UC Santa Barbara, USA. DOI 10.1109/CBD.2017.

TALLAT, Raiha; LATIF, Rana M. Amir; ALI, Ghazanfar; ZAHEER, Ahmad Nawaz; FARHAN, Muhammad; SHAH, Syed Umair Aslam. **Visualization and Analytics of Biological Data by Using Different Tools and Techniques.** 16th International Bhurban Conference on Applied Sciences & Technology. Department of Computer Science GCUF Sahiwal Campus; Department of Computer Science COMSATS University Islamabad; Department of Management Sciences GCUF Sahiwal Campus, 2019.

TANEJA, Kunal; ZHU, Qian; DUGGAN, Desmond; TUNG, Teresa. **Linked Enterprise Data Model and its use in Real Time Analytics and Context-Driven Data Discovery.** IEEE International Conference on Mobile Services. Accenture Technology Labs, California State. DOI 10.1109/MS.2015

YU, Bin; ZHANG, Chen; LI, Yitong; SUN, Jiangyan. **Research and Implementation of Data Fusion Method Based on RDF.** IEEE 3rd International Conference on Big Data Analysis. School of Computer Science and Technology, Xidian University; School of Engineering Xi'an International University. 978-1-5386-4794-3/18/\$31.00 2018 IEEE